Para el proceso del desarrollo del ETL.

En primera instancia se realizó todo lo relacionado a los puntos de extracción de datos, para esto se realizó el bucket en GCP y se cargaron todos los .csv dados por el enunciado.

Posteriormente se realizó el BigQuery, con esto ya se tendría una base de datos que se encargue del manejo de la información.

Con esto hecho se realizó la carga del flujo (el que venía incluido en el enunciado). Una vez cargado el flujo, se hace la modificación para la extracción de los datos. Los datasets se asociaron con el bucket respectivo al archivo asociado a cada entidad.

Una vez hecha la extracción a cada output se le asocia con el BigQuery para poder crear las tablas y todo lo relacionado al SQL. Esto lo que hace es crear las dimensiones de Fechas, OrderLines y StockItems. Nótese que el OrderLines corresponde a la tabla de hechos.

Ya terminado esto se realizó la extracción, la transformación y la carga de datos para el packageType, esto teniendo en cuenta las dimensiones ya generadas como una guía.

En materia de los recipies, el flujo de profesores contenía ya algunos de los cuáles se les realizó el respectivo análisis.

En el caso de Orders, lo primero que se hizo es extraer todo lo relacionado a las fechas, que es día, mes y año, se eliminan los duplicados y dada la unicidad, se genera una ID este conjunto incluyendo al ID es el que genera la dimensión de Date. Sin embargo, para unificarlo a la tabla de hechos se agarran las llaves primarias generadas a partir de la unicidad con el order_date (esto mediante un Join) y se obtiene la tabla para el conjunto de hechos.

La recipe de OrderLinesParametrizado filtra las ordenes en la tabla de hechos.

Lo mismo con el StockItems, con reservados cambios. Este escoge herramientas importantes, que esas herramientas importantes siendo estas ID, nombre, color, tamaño y precio final. Una vez seleccionados, genera un ID. Con este ID se crea la tabla de hechos y la dimensión respectiva.

Por último, para el caso de packageType se identificó que la columna LastEditedBy era irrelevante en términos del negocio, por lo que se eliminó, con la ID ya predefinida se generó la agrupación a la tabla de hechos; ya con esto además se generó el output correspondiente en el BigQuery